Basic course on regression modelling

Mervi Eerola

Turku Center of Statistics

24-25.9 and 1.10.2015

Outline of the course

1. Basic principles of statistical inference
2. Linear regression: the basics
3. Linear regression: modifying the model
4. Logistic regression

Gelman, A. & Hill, J. (2007): Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge University Press.

1. Basics of statistical inference and modelling

"Statistics concerns what can be learned from data."

Basic concepts of 'statistical language'?

1. How to deal with randomness? Random variables
2. Systematic behaviour of randomness? Probability distributions
3. Why is 'big n' better than 'small n'? The Central Limit Theorem
4. Knowing the uncertainty of estimating two means, what can we say about the uncertainty of estimating the *difference* of the two means? Transformation of statistics
5. Statistical inference? Estimating unknowns and testing hypotheses under uncertainty

Random variables

- Random variables represent variability in the outcome of interest: heights of men and women in population, earnings, education level, voting behaviour etc.
- Randomness in a variable corresponds to the fact that we cannot predict its value deterministically
- The *expectation* of a random variable is its most probable value

Random variables (cont.)

- **Example:** Let Y be a random variable telling the sex of a randomly selected newborn child
- We know *a priori* that $P(\text{girl}) \approx 0.49$ and $P(\text{boy}) \approx 0.51$
- If $Y =$"the sex of a newborn", then the expectation $E(Y)$ is "a boy", the most probable outcome
- However, the expectation of Y is $E(Y) = 0.49$ if we define
  - Y="the sex of a newborn is girl"$= 1$
  - Y="the sex of a newborn is boy"$= 0$
- Expectation is a weighted sum of possible values

$$E(Y) = P(Y = 0) \cdot 0 + P(Y = 1) \cdot 1 = 0.49$$

Probability distributions

- Probability distributions characterize variability of random variables and are building blocks for statistical models.
- Probability distributions appear in
  - *distributions of data*
  - *distributions of unknown parameters*
  - *distributions of error terms*

Probability distributions in statistical models

- In statistical modelling
  - probability distributions are used to fit a distribution to the data (say, y),
  - then using predictors $X = (x_1, x_2, ..., x_p)$,
  - model the outcome y given predictors X with some errors ϵ.
- In the model, the variation of outcome y is split to systematic (=explained by the variation in X) and random variation
- Information in the predictors X usually reduces random variation and thereby changes the distribution of the errors

Parameters of probability distributions and data

- The expected value or mean $\mu = E(y)$ of a random variable is estimated from data of size n as the *mean* and is denoted by "hat"

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Variation around the expected value, the *variance* σ^2, is estimated from data as

$$\hat{\sigma}_y^2 = s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- *Standard deviation* measures the magnitude of variation around the mean value

$$sd(y) = s_y = \sqrt{\hat{\sigma}_y^2}$$

An important result: The Central Limit Theorem

- The sum of many small independent random variables is a random variable with an approximatively normal distribution
- If $z = \sum_{i=1}^{n} z_i$ is the sum of independent components z_i, the mean μ of z is the sum of means $\mu = \sum_{i=1}^{n} \mu_i$, and its variance is the sum of variances $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$
- Write $z \sim N(\mu, \sigma^2)$: "z follows a normal distribution with expectation μ and variance σ^2"

An important result: The Central Limit Theorem

- The Central Limit Theorem holds in practice: e.g the observed heights of Finnish women actually follow a normal distribution
- BUT the height of Finnish men and women together is not so closely normal because the distributions have different means and variances

Linear transformation of normal variables are normal

- Example: If men's heights in inches are normal with mean 69.1 in and standard deviation (sd) 2.9 in then their heights in cm's are still normal with mean 175 cm and sd 7.4 cm ($y_{cm} = 2.54 \times y_{in}$)
- The *difference* of the average heights of 100 men and 100 women in a random sample is still normal with mean

$$\hat{\mu} = \bar{y} = 175 - 162 = 13cm$$

and standard deviation

$$sd(\hat{\mu}) = \sqrt{7.4^2/100 + 6.9^2/100} = 1.01$$

Linear combination of normal variables is normal

- If x and y are normal random variables with means μ_x, μ_y and variances σ_x^2, σ_y^2 and *correlation* ρ, their *sum* has mean

$$\mu = \mu_x + \mu_y \quad \text{and} \quad sd(\mu) = \sqrt{\sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y}$$

- More generally, a *linear combination* or a *weighted sum* $ax + by$ is normal with mean

$$\mu = a\mu_x + b\mu_y$$

and standard deviation

$$sd(\mu) = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y}$$

- How about the *difference* $\mu = \mu_x - \mu_y$?

Regression coefficients are approximately normal

- The unknown parameters of a regression model are its regression coefficients
- Regression coefficients are *linear combinations of data* so by The Central Limit Theorem, in a large enough sample, they are approximately normally distributed
- The same applies to regression coefficients of the logistic model and in general, to the so called maximum likelihood estimates
- This general result gives us a way of determining the uncertainty in our model results

Other probability distributions: log-normal

- All-positive continuous random variables are modelled on a logarithmic scale and follow a log-normal distribution
- Example: The logarithms of men's weights in pounds have an approximately normal distribution with mean 5.13 and standard deviation 0.17
- The mean and standard deviation of the log weight and log standard deviation are called *geometric mean* and *geometric standard deviation*
- Transformation to the original scale:

$$\hat{\mu} = \bar{y} = \ exp(\mu + \frac{1}{2}\sigma^2) = 171\text{pounds}$$

sd(y)= exp$(\mu + \frac{1}{2}\sigma^2)\sqrt{exp(\sigma^2 - 1)} = 29$ pounds

Other probability distributions: binomial

- If you have 20 shots in basket ball and each has 0.3 probability of succeeding, and the shots are independent of each other, then the number of succeeding shots has a binomial distribution with $n = 20$ and probability $p = 0.3$
- Write $y \sim Bin(n, p)$, where n is number of trials and p the expectation of $P(y = 1) =$"success"
- The model is only an approximation, in general the trials may not be independent and the probability of success p may vary from trial to trial
- The binomial model is a good starting point for a study design where the number of 'Yes' cases in a fixed number of 'trials' is of interest

Other probability distributions: Poisson

- Poisson distribution is used to model the number of 'hits' or cases in a large, often unknown, number of 'trials'
- Example: The size of Finnish population is 5.4 million and the average rate of stroke is 2364 cases per million person years. Then the annual number of strokes is modelled as Poisson with expectation of 13000 cases
- Example: In a city of 10 000 inhabitants where 0.5% of the population is named Virtanen, and the inhabitants can be taken as 'independent', the number of Virtanens in that city is a Poisson distributed variable with expectation 50
- The simple Poisson distribution is a again a starting point for more complicated models

Statistical inference: two interpretations for model

- **Sampling model:** Aims to learn some characteristic of the population which must be estimated from a *sample* or a subset of the population
- Examples: Mean earnings and standard deviations of men and women in Finland, proportions of voters for right, left and center in the population
- **Measurement error model:** Aims to learn some aspects of an underlying pattern or law but the data is measured with *error*
- Examples: Parameters α, β in the model $y = \alpha + \beta x$ are estimated from the model $y = \alpha + \beta x + \epsilon$ with error ϵ

Statistical inference: two interpretations of model

- The idea of a measurement error model apply also when complete data are observed (there is no superpopulation)
- Often the two models are combined; the sampling interpretation is that the model errors are from a population distribution $N(0, \sigma^2)$
- Example: In a random sample of students, predicting student's grade from pre-test scores only partly measures student's ability
- Note. Sometimes even the predictor variable x is measured with error (latent variable models)

Statistical inference: standard error

- The goal of statistical inference is to evaluate the *uncertainty* in the estimated parameters
- *Standard errors* are standard deviations of the parameter estimates and represent uncertainty in estimation
- Standard errors are used to construct confidence intervals for the estimates
- Standard error of *mean*: $s.e(\hat{\mu}) = \sigma/\sqrt{n}$ in a sample of size n where σ is the standard deviation of the observations
- Standard error of *proportion*: In a sample of size n, the estimated proportion of $y =$'Yes' is $\hat{p} = y/n$ and so $s.e(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$

Statistical inference: confidence interval

In large samples the confidence intervals (CI) are based on the normal distribution

- The 95% confidence interval: estimate ± 2 standard error
- The 68% confidence interval: estimate ± 1 standard error
- The 50% confidence interval: estimate $\pm 2/3$ standard error
- Interpretation: With 95% confidence the true parameter value is included in the 95% confidence interval
- In small samples, the t-distribution is used

Confidence interval: examples with R program

1. Continuous data: 95% CI for average height of students in the class (small sample)

```
n=length(y)
estimate=mean(y)
se=sd(y)/sqrt(n)
int.95=estimate + qt(c(.025,.975),n-1)*se
# quantiles of t-distribution
```

2. Proportion: 95% CI for proportion of supporters of death penalty in a survey (large sample)

```
estimate=y/n
se=sqrt(estimate*(1-estimate)/n)
int.95=estimate + qnorm(c(.025,.975))*se
# quantiles of normal distribution
```

Confidence intervals of linear transformations

- To get confidence intervals for a linear transformation of a basic parameter, transform directly the confidence intervals of the basic parameter
- Example (Number of pets): Assume that the 95% CI for the number of pets *per adult* is [0.52,0.62].
- What is the 95% CI for the *total* number of pets in a city of 100 000 inhabitants?

Confidence intervals of weighted averages

- *Example*: Suppose separate surveys on having pets were conducted in 28 European countries
- The estimated proportion of pets is a *weighted sum* of countrywise proportions

$$\hat{p}_{tot} = \frac{n_1}{n}\hat{p}_1 + \frac{n_2}{n}\hat{p}_2 + ... + \frac{n_{28}}{n}\hat{p}_{28}$$

  where n_i/n is the proportion of adults, $\hat{p}_i$ the proportion of having pets in country i and n the total number of adults

- Standard error of the total proportion:

$$se(\hat{p}_{tot}) = \sqrt{(\frac{n_1}{n}se(\hat{p}_1))^2 + (\frac{n_2}{n}se(\hat{p}_2))^2 + ... + (\frac{n_{28}}{n}se(\hat{p}_{28}))^2}$$

Using simulation for other functions of estimates

- Confidence intervals for more complicated functions than sums, differences, averages and linear transformations can be calculated by simulation
- Example (Death penalty): A survey of 1100 persons, of whom 700 support death penalty, 300 oppose, and 100 are neutral.
- Assume that 500 were men and 500 women, of whom 75% of men and 65% of women were supporters.
- The estimated men/women ratio is 0.75/0.65=1.15 so men support death penalty 15% more than women
- What is the standard error and thus the confidence interval for the male/female *ratio* of supporters?

Using simulation for other functions of estimates

Create 10000 simulation draws in R for both men and women from known distributions

```
n=500
p.hat.men=0.75
se.men=sqrt(p.hat.men*(1-p.hat.men)/n.men)

n=500
p.hat.women=0.65
se.women=sqrt(p.hat.women*(1-p.hat.women)/n.women)

n.sims=10000
p.men=rnorm(n.sims,p.hat.men,se.men)
p.women=rnorm(n.sims,p.hat.women,se.women)
ratio=p.men/p.women
int.95=quantile(ratio,c(0.025,0.975)
```

which yields a 95% CI [1.06,1.25] for the men/women ratio

Statistical testing

- Statistical testing compares alternatives to *reject* or *not reject* a hypothesis, never to accept a hypothesis
- In fact confidence intervals already include the test: testing the hypothesis that a coefficient takes the value 0 (or any other value) corresponds to having that value in the 95% confidence interval. Then the hypothesis is not rejected at the 5% level
- Testing whether two parameters equal corresponds to testing whether the value 0 is included in the 95% confidence interval of the *difference* of the parameters
- Testing whether the parameter has only *positive* values corresponds to checking whether both ends of the 95% confidence interval are on the positive side

How to deal with statistical significance?

- A common mistake is to report only results that have statistical significance
- Note: Statistical significance is not the same as practical significance of a result
- In a very large sample almost any difference becomes statistically significant but has no practical significance
- Example: Assume that the predictive effect of height per 10cm on earnings is 10e per month with $se = 0.2$, which is statistically significant. On the other hand, an effect of 1000e with a $se = 1000$ is not statistically significant but certainly has practical value

How to deal with statistical significance?

- A more important point is that large differences in significance levels may not indicate large differences in effect estimates
- Example: Consider two independent studies with effect estimates (e.g mean) and their standard errors 25 ± 10 and 10 ± 10
- The first result is significant on 1% level whereas the latter is not statistically significant
- However, their estimated difference in the two studies is 25-10=15 and has standard error $\sqrt{10^2 + 10^2} = 14$ which is not statistically significant
- Comparison of results should always account for uncertainty which may differ

2. Linear regression: the basics

1. One predictor and multiple predictors
2. Interactions
3. Statistical inference
4. Graphical displays of data and fitted model
5. Assumptions and diagnostics

Linear regression: basic principles

- Linear regression is a method to summarize how the *average values* of a numerical outcome vary over subpopulations defined by linear functions of predictors x
- We have already seen that simple probability models (e.g normal, binomial, Poisson) are the building blocks for more complicated models
- Regression can be used to *predict* an outcome (its expectation) given a linear function of the predictors
- Multilevel models extend the setting of simple regression models (not discussed in this course)
- We start with examples and interpretations, then give an overview of the statistical theory

Example: Child's test score

- A subsample of the US National Longitudinal Survey of Youth
- Outcome of interest: Cognitive test score of 3-4 year old children given characters of their mothers
- Binary predictor: Mother graduated from high school (yes/no)
- *Model*: $\text{child.score} = 78 + 12 \cdot \text{mother.hs} + \text{error}$
- The *deterministic* part of the model excluding prediction error

$$\widehat{\text{child.score}} = 78 + 12 \cdot \text{mother.hs}$$

- This is the child's predicted or expected score given mother's graduation.
- In statistical language this reads

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Example: Child's test score

- The model summarizes *the difference in average test scores* between children in the two groups (high school/no high school)
- The *intercept* $\beta_0 = 78$ is the average or predicted score for children whose mother did *not* complete high school (`mother.hs=0`)
- The average or predicted score for children whose mother *did* complete high school is $78 + 12 \cdot 1 = 91$ so the average predicted benefit of mother's graduation is $\beta_1 = 12$ points on child's scoring

Example: Child's test score

- Continuous predictor: Mother's score on an IQ test
- *Model:* child.score $= 26 + 0.6 \cdot$ mother.IQ $+$ error
- Interpretation: In sub-populations that differ in maternal IQ with one point (or 10 points), the child's expected test score is on average 0.6 (or 6 points) higher than in the group of lower maternal IQ.
- Note: Interpretation of the intercept is problematic in this case: `mother.iq=0`!

Linear regression: multiple predictors

- Interpretation of regression coefficients becomes more complicated when several predictors are included
- In general, all effects depend on other predictors in the model
- Usual assumption: *With all other predictors held constant*, change in one unit in the predictor of interest is the value of the regression coefficient
- This implies that we compare the regression slopes of individuals who differ in one predictor while being at the same level of the other predictors
- This may not always be possible: In the Child test score example, testing e.g the quadratic effect of IQ needs both IQ and IQ^2 in the model

Linear regression: two interpretations

- **Predictive interpretation:** Average difference in the outcome variable when changing one predictor x by 1 unit to $x + 1$ (in the group of individuals having x) and being identical in all the other predictors (group level)
- **Counterfactual interpretation:** Expected change in the outcome variable caused by adding 1 to one predictor while leaving all the other predictors unchanged (individual level)
- If the mother's IQ *were* 101 instead of 100, which it is, the expected increase in the child's score *would be* 0.6
- Note. Counterfactual interpretation is *causal* and requires additional assumptions about comparable subgroups

Linear regression: interactions

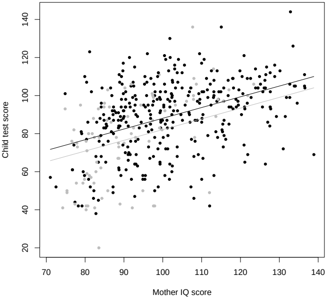
- In the previous model the regression slope of mother's IQ was forced to equal across subgroups of high school completion. What if this is not the case?
- A model with interaction term `mother.hs`×`mother.IQ` allows the regression slope to vary across subgroups
- Model:

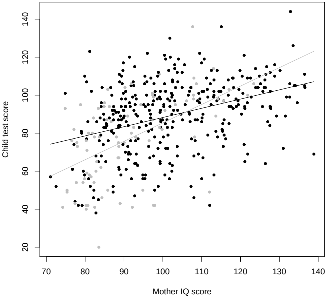
$$\begin{aligned}\text{child.score} = & -11 + 1.1\text{mother.IQ} + 51 \cdot \text{mother.hs} \\ & -0.5 \cdot \text{mother.hs} \times \text{mother.IQ} + \text{error}\end{aligned}$$

is a combination of two lines:

$$\begin{aligned}-11 + 1.1\text{mother.IQ} + 51 \cdot 1 - 0.5 \cdot \text{mother.hs} \times \text{mother.IQ} \\ -11 + 1.1 \cdot \text{mother.IQ}\end{aligned}$$

Linear regression: interactions





Linear regression: interpretation of coefficients

- `Intercept:` predicted test score for children whose mother did not complete high school and whose IQ=0 (not meaningful)
- `mother.hs`: difference between the predicted test scores for children whose mother did not complete high school and whose IQ=0 and children whose mother did complete high school and whose IQ=0 (not meaningful)
- `mother.IQ`: difference between the predicted test scores for children whose whose mother did not complete high school but whose mothers differ by 1 in IQ (meaningful)
- Interaction `mother.hs` × `mother.IQ`: difference in the slopes of `mother.IQ` in the subgroups of mother completed/not completed high school (meaningful)